Exploring Evidence Aggregation Methods and External Expansion Sources for Medical Record Search

Dongqing Zhu and Ben Carterette
Dept. of Computer & Information Sciences
University of Delaware, Newark, DE USA

1 Introduction

This paper describes and analyzes experiments we performed for the Medical Records track in the 2012 Text REtrieval Conference (TREC). We mainly investigated three research problems:

1. Evidence Aggregation: In last year's track there were two different methods in general for obtaining a visit ranking out of reports (smaller document units), i.e., (A) using reports as indexing and retrieval units and then converting a report ranking into a visit ranking, and (B) using visits as indexing and retrieval units by concatenating reports at the very first stage and then obtain a visit ranking directly.

   Method A avoids the potential problem of varying visit document length, while Method B naturally aggregates evidence scatter over multiple reports and easily obtains a visit ranking. It is unclear which method is better based on all reported results. Thus, we compared the two approaches, tried various score aggregation methods for (A), and combined both approaches in a way that further improved the system performance.

2. Expansion Sources: We tested a variety of external collections (ranging from general web datasets to domain-specific thesauri, and from Megabyte datasets to Terabyte datasets) for query expansion, compared their effectiveness, and obtained useful insights into the data.

3. Retrieval Models: We tested several statistical IR models (proven to be effective on news and web collections) on this medical collection, and explored ways to combine these models to address different aspects of task. For instance, we used MRF model to model term proximity since most medical concepts are phrases. We also used a mixture of relevance models to obtain various relevant expansion terms covered by different expansion collections respectively, which is expect to significantly alleviate the vocabulary mismatch between medical terminologies.

For TREC submissions, we tested systems that combined multiple IR models, leveraged diverse expansion sources, and used various evidence aggregation methods. We implemented all the retrieval models in the Indri[1] retrieval system.

[1] http://www.lemurproject.org/indri/

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Exploring Evidence Aggregation Methods and External Expansion Sources for Medical Record Search	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Delaware,Dept. of Computer & Information Sciences,Newark,DE,19716**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

2 System Features

In this section, we describe the retrieval models and evidence aggregation strategies.

2.1 Retrieval Model

In this section, we briefly describe the three retrieval models, namely the query likelihood language model, MRF model, and mixture of relevance models, which serve as the underpinning of our retrieval systems.

Baseline Model
Our baseline model is the query likelihood language model as formulated below:

$$\text{score}(D, Q) = \log P(Q|D) = \sum_{i=1}^{n} \log \frac{tf_{q_i,D} + \mu \frac{tf_{q_i,C}}{|C|}}{|D| + \mu}, \tag{1}$$

where q_i is the ith term in query Q, n is the total number of terms in Q, $|D|$ and $|C|$ are the document and collection lengths in words respectively, $tf_{q_i,D}$ and $tf_{q_i,C}$ are the document and collection term frequencies of q_i respectively, and μ is the Dirichlet smoothing parameter.

Markov Random Field Model
Medical queries usually contain phases that describe conditions, symptoms, drug names, treatments, etc. These query terms are likely to occur in close proximity to each other in relevant documents. Thus, we use the Markov random field (MRF) model proposed by Metzler and Croft [9] to model term dependencies. We use their sequential dependence model in particular. Following Metzler and Croft [9], we set the feature weights $(\lambda_T, \lambda_O, \lambda_U)$ to (0.8, 0.1, 0.1).

Mixture of Relevance Models
Queries specified by the search users can have a "vocabulary mismatch" with the content in a medical report since there are many different ways to express a medical concept (e.g., "hearing loss", "hearing impairment", "difficult of hearing", and even "deafness" are all semantically related, but they only have one common term at the most). The consequence is that the system may have a relatively low recall if there is a "vocabulary mismatch". We can alleviate this problem and improve our baseline retrieval model by expanding the query with additional "related" terms. These related terms (also called expansion terms) can be derived from a relevance model θ_Q, which is usually built upon top-ranked k documents for the query in the target collection (i.e., the same collection used for retrieval).

Thus, in this paper we derive expansion terms based on their weights p which are estimated by:

$$p_i = \sum_{j=1}^{k} \exp\{\frac{tf_{e_i,D_j}}{|D_j|} + \log \frac{|C|}{df_{e_i,C}} + score(D_j, Q)\}, \tag{2}$$

where $score(D_j, Q)$ is the query likelihood score for the top jth feedback document in the initial retrieval set ranked by Equation 1, tf_{e_i,D_j} is the term frequency of e_i in document D_j, $df_{e_i,C}$ is the document frequency of e_i in collection C, and $|D_j|$ and $|C|$ are document and collection lengths in words respectively. This formula estimates the importance of term e_i based on its term frequency, inverse document frequency, and feedback document scores. m terms with highest scores p are selected as expansion terms, and they form our estimated relevance model $\hat{\theta}_Q$. Note that we also normalize p so that we have an estimated probability $P(w|\hat{\theta}_Q)$ for each word w.

Relevance modeling can be further improved upon by leveraging information in other document collections. Specifically, following Diaz and Metzler [2], we can form relevance models for two or more additional collections, then expand the query using those models.

To achieve better performance, we linearly interpolate the mixture of relevance models with the maximum likelihood (ML) query estimate by formulating the equation:

$$P(w|\theta_Q) = \lambda_Q \frac{\#(w, Q)}{|Q|} + \sum_{C} \lambda_C P(w|\hat{\theta}_{Q,C}), \tag{3}$$

where the first part is the weighted ML query estimate for word w and the second part represents the mixture of relevance models. In particular, $P(w|\hat{\theta}_{Q,C})$ is the probability of w in the estimated relevance model $\hat{\theta}$ built upon top-ranked documents in expansion collection C. λ's are collection weights and $\lambda_Q + \sum_C \lambda_C = 1$. For TREC submissions, we set λ's to (0.7, 0.1, 0.1, 0.1) and use top 10 terms from top 50 feedback documents. We implement Equation 3 using Indri in the same way as our previous work [12]. We denoted this model as MRM.

A Combined Model

We linearly combine MRF and MRM to get our third retrieval model. The scoring function looks like

$$P(w|\theta_Q) = \lambda_Q \cdot \text{MRF} + \sum_C \lambda_C P(w|\hat{\theta}_{Q,C}), \tag{4}$$

which is similar to Equation 3. The difference is that we replace the ML query estimate with MRF. The new retrieval model is expected to benefit from term dependence modeling as well as query expansion.

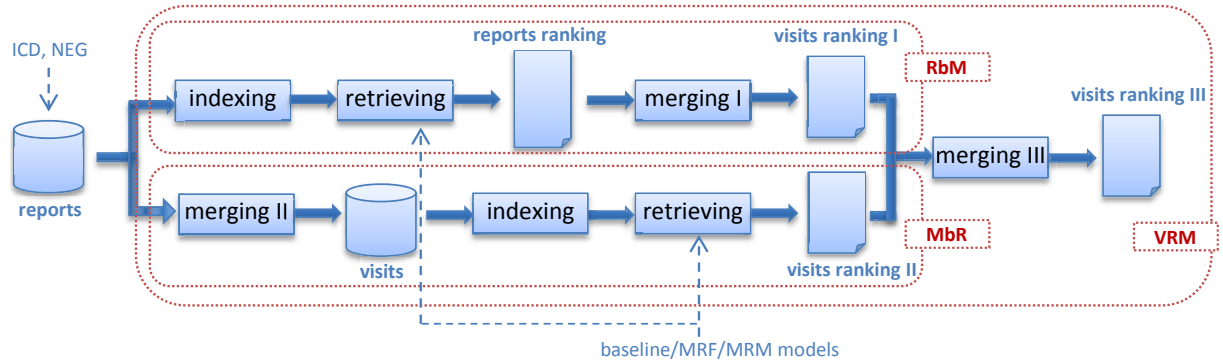


Figure 1: Merging results from two different retrieval methods.

2.2 Multi-level Evidence

In the aforementioned retrieval models, "document" (i.e., D) is a broad term that could indicate different granularities of a visit: the text of all reports associated with the visit, a single report from the visit, or just one field within a single report from the visit. Thus, in this section we describe how we leverage evidence from each of these to come up with a final document score.

Field Level Evidence

The main fields in a report are the doctor's notes and the fields that contain diagnosis codes. Here we describe how we leverage ICD-9 codes in the language model, and how we remove some extraneous information and extract useful information from doctor's notes.

Code Expansion: We expand ICD codes in the "admit diagnosis" and "discharge diagnosis" fields with their corresponding descriptions[2] to introduce additional useful words into the documents. We refer this feature as ICD in the following sections.

Negation Removal: The "report text" filed contains clinical narratives. One distinct feature of clinical narratives is that negation phrases are frequently used to claim the absence of certain conditions or symptoms [1], such as "cannot tell", "not clear", "without evidence", etc. Negations may cause retrieval false positives. Thus, we use NegEx[3] [4], an open-source clinical negation detection tool, to remove all negated portions of the sentences from the medical records before indexing. We refer this feature as NEG in the following sections.

Age/Gender Filtering: We use simple regular expressions to search for age/gender indication words and phrases in both the"report text" filed and the topics. We use the extracted age and gender information

[2] https://drchrono.com/public_billing_code_search
[3] http://code.google.com/p/negex/

to filter from the retrieval set visits that do not meet the inclusion criteria specified in the topics. We refer this feature as AGF in the following sections.

Previous work [7, 8, 3, 12] have shown that the above three medical features are quite useful. Thus, they would be the default features for our systems unless otherwise specified.

Report Level Evidence
Evidence in a visit may mainly exist in only a small proportion of all the associated reports. This allows us to rely on the strongest evidence of a visit to estimate its relevance. Thus, we use reports as the initial retrieval units (i.e., building an index for reports and applying the retrieval model to each report), and then transform a report ranking into a visit ranking based on the strongest report-level evidence, which is equivalent to using the following report score merging method for ranking visits:

$$\text{score}_{\text{RbM}}(V, Q) = f_{\text{RbM}}(\{\text{score}(r_1^V, Q), \text{score}(r_2^V, Q), ...\}), \tag{5}$$

where r_j^V is a report associated with visit V based on the report-to-visit mapping, $\text{score}(r_j^V, Q)$ is the language modeling score of the report with respect to query Q, and f_{RbM} is the function for aggregating the scores. We will try MAX, SUM, and ANZ for f_{RbM} in Section 3. We name this evidence aggregation strategy Retrieval-before-Merging (RbM). The merging process involved in RbM corresponds to "merging I" in Figure 1.

Visit Level Evidence
Evidence may also spread across multiple reports, especially when the information need is a complex one. Thus, our second strategy to aggregate evidence is to first merge reports from a single visit field by field into a visit document and then construct an index for visit documents. With this strategy, the language model built on a merged document can naturally combine the evidence scattered across multiple reports. Furthermore, this strategy can directly lead to a ranking of visits which are the desired retrieval units. We call this second evidence aggregation strategy Merging-before-Retrieval (MbR). The merging process involved in MbR corresponds to "merging II" in Figure 1.

Top-level Evidence
RbM and MbR as described above are two different strategies for aggregating evidence and ranking visits. RbM and MbR complement each other in that the former can naturally aggregate evidence spreading across multiple reports (which would be challenging to do at the report-level) while the latter can leverage the strongest evidence (which may become less apparent after reports merging in MbR) to estimate relevance. This leads to our third evidence aggregation method in which we take advantage of both RbM and MbR by merging their visit rankings, as demonstrated by "merging III" in Figure 1. We call third strategy as Visit-Ranking-Merging (VRM). The merging method (i.e., "Merging III" in Figure 1) is defined by:

$$\text{score}_{\text{VRM}}(V, Q) = f_{\text{VRM}}(\text{score}_{\text{RbM}}(V, Q), \text{score}_{\text{MbR}}(V, Q)), \tag{6}$$

where $\text{score}_{\text{RbM}}(V)$ and $\text{score}_{\text{MbR}}(V)$ are the language modeling scores for visit V with respect to query Q in the two visit rankings obtained by RbM and MbR respectively, f_{VRM} is the function for score aggregation, and $\text{score}_{\text{VRM}}(V, Q)$ is the final score of visit V in the merged ranking. We will try different methods for f_{VRM} such as CombMNZ, CombSUM, and CombMAX in Section 3 below.

2.3 Experimental Setup

We explore the best system feature settings for 2012 TREC submission by performing cross-validation on the test collection of 2011 medical records track. Once we find the best settings as will be described in Section 3, we train the systems on the whole 2011 test collection and generate runs for 2012 topics.

We use the Indri retrieval system for indexing and retrieving. In particular, we use the Porter stemmer to stem words in both reports and queries, and use a simple standard medical stoplist [5] for stopping words in queries only. Then we conduct 5-fold cross-validation and use top 1000 retrieved visits for each query to evaluate our system under different settings. In each iteration, we train our system on 28 queries to obtain the best parameter setting for MAP by sweeping over the range of [1000, 20000] at a step size of 1000 for the Dirichlet smoothing parameter (i.e., μ in Equation 1), and then generate a ranking for each of the remaining

7 queries based on the trained system. When complete, we have full rankings for all 35 topics as a test set. We evaluate the system based on the MAP, bpref, P10, and Rprec over all 35 topics.

We train our systems on MAP though bpref is the primary evaluation metric for 2011 medical track. There are two reasons: 1) training on MAP is most commonly used in IR to improve retrieval performance; 2) we find that training on MAP improves the retrieval performance on other metrics as well while training on bpref does not improve the overall performance. Thus, MAP and bpref will both be the primary evaluation measures in this work. In fact, MAP correlates well with bpref as we will show in the next section.

To access the statistical significance of differences in the performance of two systems, we perform one-tailed paired t-test for MAP (since we train systems on MAP).

3 Experiments and Results

This section describes experiments, presents the evaluation results, and discusses the research findings.

3.1 Retrieval before Merging

As mentioned in Section 2.2, we have several options for choosing the score merging function f_{RbM} in Equation 5 (i.e., "merging I" in Figure 1) for RbM. Now we describe them formally below:

MAX:

$$\text{score}_{\text{RbM}}(V, Q) = \max(\{\text{score}(r_j^V, Q)\})$$

SUM:

$$\text{score}_{\text{RbM}}(V, Q) = \sum_j \text{score}(r_j^V, Q)$$

ANZ:

$$\text{score}_{\text{RbM}}(V, Q) = \frac{\sum_j \text{score}(r_j^V, Q)}{|\{\text{score}(r_j^V, Q) \neq 0\}|}$$

where again $\text{score}(r_j^V, Q)$ is the language modeling score of the report r_j^V (associated with visit V) with respect to query Q. ANZ stands for "Averaging over Non-Zeros", meaning we only consider reports containing at least one query term. MAX, SUM, and ANZ are similar to CombMAX, CombSUM, and CombANZ proposed by Fox and Shaw [11]. However, CombMAX, CombSUM and CombANZ were used for merging multiple retrieval runs.

Table 1 shows that MAX is superior to SUM and ANZ. This confirms our assumption that we can rely on the strongest evidence (i.e, the most relevant report) of a visit to estimate the relevance of that visit. Thus, we will use MAX for score merging in RbM by default in this paper.

	MAX (selected)	SUM	ANZ
MAP	0.416	0.110	0.317

Table 1: Comparison of score merging methods for RbM.

3.2 Evidence Aggregation

Similarly, we also have several options for choosing the score merging function f_{VRM} in Equation 6 (i.e.," merging III" in Figure 1) for VRM, such as CombMNZ, CombANZ, and CombMAX [11]. In our case, we are only merging two rankings. Thus, these merging methods are specified as follows:

CombMNZ:

$$\text{score}_{\text{VRM}}(V, Q) = N_V \cdot [\text{score}_{\text{RbM}}(V, Q) + \text{score}_{\text{MbR}}(V, Q)]$$

CombSUM:

$$\text{score}_{\text{VRM}}(V, Q) = \text{score}_{\text{RbM}}(V, Q) + \text{score}_{\text{MbR}}(V, Q)$$

CombMAX:

$$\text{score}_{\text{VRM}}(V,Q) = \max(\text{score}_{\text{RbM}}(V,Q), \text{score}_{\text{MbR}}(V,Q))$$

CombANZ:

$$\text{score}_{\text{VRM}}(V,Q) = \frac{\text{score}_{\text{RbM}}(V,Q) + \text{score}_{\text{MbR}}(V,Q)}{N_V}$$

where $\text{score}_{\text{VRM}}(V,Q)$ is the merged score for visit V, and $\text{score}_{\text{RbM}}(V,Q)$ and $\text{score}_{\text{MbR}}(V,Q)$ are the scores for V in two different visit rankings as demonstrated in Figure 1, and N_V is the number of rankings that have V in the top 1200 retrieved visits (We cut off the merged rank list at rank 1000 to get the final ranking). Note that $\text{score}_{\text{MbR/RbM}}(V,Q) = 0$ if V does not appear in the top 1200 retrieved. We compare the performance of these merging methods using the primary evaluation measures in Table 2. As we can see, CombMNZ and CombSUM achieve comparable performance, and are better than CombMAX and CombANZ. Thus, we can infer that a good aggregation strategy for "merge III" should favor visits that appear in both rankings. We use CombMAX and CombANZ as the merging methods for VRM.

Method	**MAP**	**bpref**
CombMNZ (selected)	0.446	0.564
CombSUM (selected)	0.446	0.563
CombMAX	0.427	0.559
CombANZ	0.356	0.510

Table 2: Comparison of score merging methods for VRM.

Next, we compare the three evidence aggregation strategies as described in Section 2.2. Table 3 shows that VRM is significantly better than MbR and RbM on MAP, which means that merging visit rankings as the top-level evidence aggregation strategy boosts the retrieval performance significantly.

System	**MAP**	**bpref**	**P10**	**Rprec**
MbR	0.393	0.530	0.565	0.403
RbM	0.416	0.551	0.594	0.434
VRM	$0.446^{\triangle}$	0.563	0.635	0.456

Table 3: Comparison of evidence aggregation methods. $^{\triangle}$ indicates that the MAP difference between VRM and MbR/RbM is statistically significant ($p < 0.05$).

3.3 Selection of Expansion Collections

We test several expansion collections. In addition to the medical records that are the target of retrieval, we leverage information in several other large, widely-available collections: ImageCLEF 2009 Medical Image Retrieval Task dataset [10], TREC 2007 Genomics Track dataset [6], TREC 2009 ClueWeb09 Category B dataset (excluding Wikipedia pages), a Wikipedia dataset (containing those excluded Wikipedia pages), and the 2012 Medical Subject Headings (MeSH). Table 4 provides detailed information about these datasets. In particular, we use MeSH for expansion in the way as described in [13]. Moreover, the CLEF dataset consists of 74,902 medical images. We crawled 5,704 full-text CLEF articles associated with these images as the actual external collection used in this work. We choose these collections because there are existing topics and relevance judgments for analysis and because we want to compare the effects of different sources on retrieval performance.

For simplicity, we use aggregation strategy MbR (without any medical features described in Section) and retrieval model MRM with one expansion collection at a time to explore the expansion effectiveness of each collection as show in Table 5.

As we can see in Table 5, ImageCLEF and Wikipedia have comparable improvement over the baseline, though the former is more medical-related, much smaller, and less noisy than the latter. The same situation applies to Genomics and ClueWeb09. However, Genomics and ClueWeb09 are much larger than

Collection	# documents	vocabulary size	avg doc length
Medical*	100,866	10^5	423
MeSH	n/a	n/a	n/a
ImageCLEF	5,704	10^5	6,495
Genomics	162,259	10^7	6,595
Wikipedia	5,957,529	10^6	1,305
ClueWeb09	44,262,894	10^7	756

Table 4: Collection Statistics.

System	MAP	Significance	bpref	P10
Baseline (B)	0.353		0.469	0.506
ImageCLEF (I)	0.371 (+5.1%)		0.492	0.544
Wikipedia (W)	0.376 (+6.5%)		0.500	0.550
ClueWeb09 (C)	0.390 (+11%)	$>\{B\}$	0.513	0.556
MeSH (S)	0.391 (+11%)	$>\{B, I\}$	0.496	0.547
Medical (M)	0.393 (+11%)	$>\{B\}$	0.520	0.535
Genomics (G)	0.395 (+12%)	$>\{B, W\}$	0.524	0.553

Table 5: Evaluation of query expansion. "$X > S$" means the MAP difference between system X and any system specified in set S is statistically significant. The statistical significance is determined using one-tailed paried t-test on queries and p-value < 0.05.

runID	Features				Scores				
	MRF	MRM		VRM	MAP	infAP	infNDCG	Rprec	P10
		Genomics+Medical+ClueWeb	MeSH						
udelSUM	√	√	√	CombSUM	0.413	0.286	0.578	0.419	0.592
udelMNZ	√	√	√	CombMNZ	0.412	0.285	0.576	0.418	0.594
udelMRF	√	√		CombMNZ	0.408	0.280	0.572	0.415	0.604
udelMED		√	√	CombMNZ	0.398	0.269	0.564	0.410	0.590

Table 6: Feature settings and results for TREC submissions.

ImageCLEF and Wikipedia respectively, and Genomics and ClueWeb09 both have significant improvement over the baseline. Genomics is also significantly better than Wikipedia. Thus, we can infer that expansion effectiveness depends on both the quality (i.e., content similarity to the target collection) and size of the expansion collection.

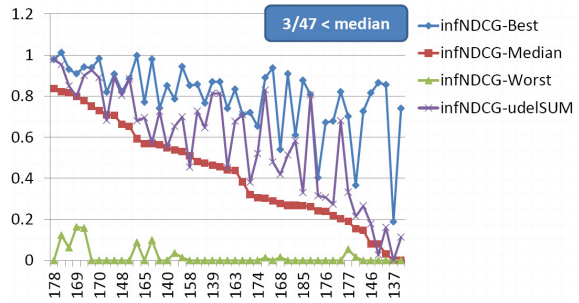
MeSH expansion is different from general expansion in that it relies on a controlled vocabulary from which expansion terms derived are not as diversified as those from a general expansion collection. For instance, for the query "hearing loss", it is difficult for MeSH to provide related expansion terms such as "cochlear", "noise", "auditory", and "binaural" (top-ranked terms from Genomics), "cerumen", "canals", and "tympanic" (from Medical), "vestibular", "ear", and "stape" (from ImageCLEF). Some of these terms do appear in the MeSH trees at upper levels, however, it is hard to find a link to them, i.e., discriminating them from other unrelated tree nodes. Simply including all visited concepts along the path is likely to cause query drift. Moreover, these terms normally appear in phrase concepts having different meanings than individual terms.

MeSH expansion is quite restrictive, yet is comparable to top performing single expansions and is significantly better than the baseline and ImageCLEF. This is most likely because our MeSH expansion emphasizes modeling term proximity which is a big advantage of any medical thesaurus-based expansion over general expansion. Another merit of MeSH expansion is that, if used properly, it rarely includes bad expansion terms, while we have no control of the quality of each expansion term from general expansions.

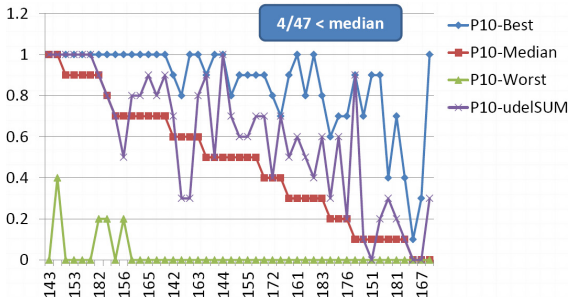
Based on Table 5, we choose Genomics, Medical, MeSH, and ClueWeb09 (i.e., the top-performing expansion collections) for our MRM model.

3.4 TREC Submissions and Results

Base on all the previous investigations, we select and combine multiple features for our TREC submissions as shown in Table 6. The settings for udelMRF and udelMED are for evaluating the impact of MRF and MeSH



(a) **udelSUM** is below TREC medians for 3/47 topics on **infNDCG**.



(b) **udelSUM** is below TREC medians for 4/47 topics on **P10**.

Figure 2: Comparison with TREC results.

	udelMNZ	udelMRF	udelMED
udelSUM	0.1635	0.0190	0.0005
udelMNZ	–	0.0335	0.0008
udelMRF	–	–	0.0181

Table 7: Pairwise one-tail paired t-test on **infAP**

respectively. Table 6 also shows the evaluation scores averaged over 47 official topics. We pick udelSUM, the system with the highest MAP score, for further analysis. Figure 2 shows the comparison of infNDCG and P10 scores with TREC results (combining both automatic and manual runs). As we can see, system udelSUM is above TREC medians for the majority of topics. We observe similar results for the other three runs.

Table 7 shows the results of pairwise one-tail paired t-test on infAP for our four submitted runs. The significance scores indicate that MRF and MeSH are both very effective system features.

4 Conclusion and Future Work

For 2012 Medical Records track, we investigated various evidence aggregation methods, explored different query expansion collections, and combined multiple statistical IR models. Our systems perform well compared with the aggregated TREC results. In particular, we found the following to be very effective: 1) external expansion using diverse sources, 2) models that incorporate term proximity information, 3) evidence aggregation at both report and visit levels. For future work, we plan to investigate why our systems did not do well on a few hard topics.

References

[1] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. Evaluation of negation phrases in narrative clinical reports. *Proceedings of AMIA Symposium*, pages 105–109, Jan. 2001.

[2] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2006. ACM.

[3] T. Goodwin, B. Rink, K. Roberts, S. M. Harabagiu, and R. Tx. Cohort shepherd: Discovering cohort traits from hospital visits. In *Proceedings of The 20th Text REtrieval Conference*, 2011.

[4] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851, 2009.

[5] W. Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Health Informatics. Springer, third edition, 2009.

[6] W. R. Hersh, A. M. Cohen, L. Ruslen, and P. M. Roberts. TREC 2007 genomics track overview. In *TREC*, 2007.

[7] B. King, L. Wang, I. Provalov, and J. Zhou. Cengage Learning at TREC 2011 medical track. In *Proceedings of The 20th Text REtrieval Conference*, 2011.

[8] N. Limsopatham, C. Macdonald, I. Ounis, G. Mcdonald, and M. Bouamrane. University of Glasgow at medical records track 2011: Experiments with Terrier. In *Proceedings of The 20th Text REtrieval Conference*, 2011.

[9] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, page 472, 2005.

[10] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C. E. Kahn, and W. R. Hersh. Overview of the CLEF 2009 medical image retrieval track. In *CLEF (2)*, pages 72–84, 2009.

[11] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

[12] D. Zhu and B. Carterette. Using multiple external collections for query expansion. In *Proceedings of The 20th Text REtrieval Conference*, 2011.

[13] D. Zhu and B. Carterette. Improving health records search using multiple query expansion collections. In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine*, 2012.